Bioinformatics, 31, 2015, i348–i356
doi: 10.1093/bioinformatics/btv259
ISMB/ECCB 2015

OXFORD

Knowledge-driven geospatial location resolution for phylogeographic models of virus migration

Davy Weissenbacher[1,*], Tasnia Tahsin[1], Rachel Beard[1,2], Mari Figaro[1,2], Robert Rivera[1], Matthew Scotch[1,2] and Graciela Gonzalez[1]

[1]Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259, USA and [2]Center for Environmental Security, Biodesign Institute, Arizona State University, Tempe, AZ 85287-5904, USA

*To whom correspondence should be addressed.

Abstract

Summary: Diseases caused by zoonotic viruses (viruses transmittable between humans and animals) are a major threat to public health throughout the world. By studying virus migration and mutation patterns, the field of phylogeography provides a valuable tool for improving their surveillance. A key component in phylogeographic analysis of zoonotic viruses involves identifying the specific locations of relevant viral sequences. This is usually accomplished by querying public databases such as GenBank and examining the geospatial metadata in the record. When sufficient detail is not available, a logical next step is for the researcher to conduct a manual survey of the corresponding published articles.

Motivation: In this article, we present a system for detection and disambiguation of locations (toponym resolution) in full-text articles to automate the retrieval of sufficient metadata. Our system has been tested on a manually annotated corpus of journal articles related to phylogeography using integrated heuristics for location disambiguation including a distance heuristic, a population heuristic and a novel heuristic utilizing knowledge obtained from GenBank metadata (i.e. a 'metadata heuristic').

Results: For detecting and disambiguating locations, our system performed best using the metadata heuristic (0.54 Precision, 0.89 Recall and 0.68 F-score). Precision reaches 0.88 when examining only the disambiguation of location names. Our error analysis showed that a noticeable increase in the accuracy of toponym resolution is possible by improving the geospatial location detection. By improving these fundamental automated tasks, our system can be a useful resource to phylogeographers that rely on geospatial metadata of GenBank sequences.

Contact: davy.weissenbacher@asu.edu

1 Introduction

Zoonotic viruses are viruses that are transmittable between animals and humans. Leading to the rise or the re-emergence of various diseases such as influenza, rabies and Ebola, these viruses are an important threat to population health and are monitored by local, national and international health organizations. To ensure effective surveillance of these viruses, it is essential to understand their origins, mutations and their geospatial transmission patterns. In recent years, phylogeography, the science that studies the geospatial lineage of species (Avise, 2000), has been applied to virus genomes to model their evolution and their diffusion among human and animal hosts. Phylogeography tracks the spread of these viruses by utilizing genetic sequence data and its corresponding metadata related to the time of sample collection and the location of the host. In particular, the geospatial metadata is vital to phylogeography since it is used to recreate the migration path of the virus.

Phylogeography often includes the use of secondary data that has been deposited in public databases by other researchers. A popular resource is the GenBank database (Benson *et al.*, 2011) which is maintained by the National Center for Biotechnology Information (NCBI). With more than 1.9 million virus sequences (as of March 2015), GenBank provides abundant information on viral sequence data. However, previous work has suggested that geospatial metadata, when it is not simply missing, can be imprecise. In their article Scotch *et al.* (2011) estimate that only 20% of GenBank records of zoonotic viruses contain detailed geospatial metadata such as a

county or a town name. Thus, some GenBank records provide generic information (such as China or USA) about where the virus was found, without mentioning the specific places within these countries. More specific information, however, may be present in the related articles, and researchers are then forced to read the paper to locate these additional pieces of geospatial metadata for a given GenBank record. Since phylogeography studies of zoonotic viruses often include hundreds of sequences in a dataset, this manual process can represent a highly time-consuming and labor-intensive process. In this article, we discuss the development and evaluation of an automated approach to retrieve geospatial metadata with finer level of granularity from full-text journal articles.

One of the major challenges to retrieving geospatial metadata from free-text is toponym resolution which associates all names of places found in a document with their corresponding geospatial locations. Toponym resolution is commonly performed in two steps. The first step, *toponym detection*, finds all occurrences of names of places in the document. The second step, *toponym disambiguation*, allocates unique coordinates (longitude and latitude) to all names found in the first step. The toponym disambiguation task is made difficult due to numerous ambiguities existing between different toponyms, like Manchester, NH USA versus Manchester, UK (Geo-Geo ambiguities) and between toponyms and other entities, such as names of people or daily life objects (Geo-NonGeo ambiguities). Our resolver uses a hybrid approach combining dictionary-based and rule-based heuristics. It first uses a built-in dictionary based on the GeoNames.org database to detect mentions of names of places in articles and then, disambiguates the names of places found using knowledge of their physical properties and contextual clues. Since, to the best of our knowledge, no standard corpus exists to date for evaluating toponym resolution on phylogeography documents, we opted to create our own gold standard corpus from full-text articles available on PubMed Central.

The main contributions of this article are (i) the description of a new gold standard corpus for toponym resolution in phylogeographic texts, (ii) a rigorous evaluation of usual heuristics for toponym disambiguation on this gold standard and (iii) the proposition of an innovative heuristic exploiting knowledge from the metadata of corresponding GenBank records to improve the automatic process of disambiguation.

Section 2 defines the task of toponym resolution in more detail and reviews the approaches proposed for solving this problem in general as well as in the biological domain. Section 3 focuses on the architecture of our toponym resolver and its heuristics. After introducing our corpus and the metrics used for our evaluation in Section 4, Section 5 discusses the results achieved by our system and analyzes its errors. Section 6 concludes our work and elaborates on further improvements.

2 Toponym resolution

2.1 Toponym resolution: state-of-the-art

The aim of toponym resolution is to find all location names mentioned in a document (detection) and to assign to each of them the unique latitude and longitude coordinates corresponding to its centroid (disambiguation). Our definition excludes all indirect mentions of places such as, '30 km north from Boston' or 'the Hong Kong strain'. Our work only handles rigid references to geographic locations through the use of proper names. Detection of the toponyms has been extensively studied in named entity recognition; location names were one of the first classes of named entities to be detected in text (Piskorski and Yangarber, 2013). Disambiguation of toponyms is a more recent task. The toponym resolvers presented in the following sections may have been evaluated on their capacities to detect *and* to disambiguate toponyms in documents (End-to-End evaluation) or they may have been evaluated only on their ability to disambiguate toponyms (Disambiguation Only). In the latter evaluation method, all location names are known by the resolver but not their precise coordinates. This evaluation method is interesting for comparing different algorithms for disambiguating toponyms, regardless of the performance of their detection methods.

2.1.1 Disambiguation driven by lexical context

Two complementary approaches have been proposed in the literature to resolve toponyms. The first approach exploits lexical contexts in documents where the toponyms appear. The lexical contexts often provide various markers which are used to disambiguate toponyms (Adams and McKenzie, 2013). In Tobin *et al.*, (2010) the authors searched for place names surrounding an ambiguous toponym within a sentence to determine the context. Their approach is based on the hypothesis that the name of a place rarely appears on its own; authors of documents often provide additional information to locate the place by referring to a broader area containing this place. For example in *Paris, Ontario*, Paris is not the capital of France but the city in Canada. The system proposed was evaluated on two different corpora, the first one composed of historical documents and the second one composed of newspapers articles, the SpatialML corpus (Mani *et al.*, 2008). The authors report average scores on both corpora for Disambiguation Only (0.81 P and 0.81 P, respectively) and a score of 69.5% of toponyms correctly found by an End-to-End system on SpatialML. A known limitation of this method is that the resolution fails if there is no adjacent toponym in the context.

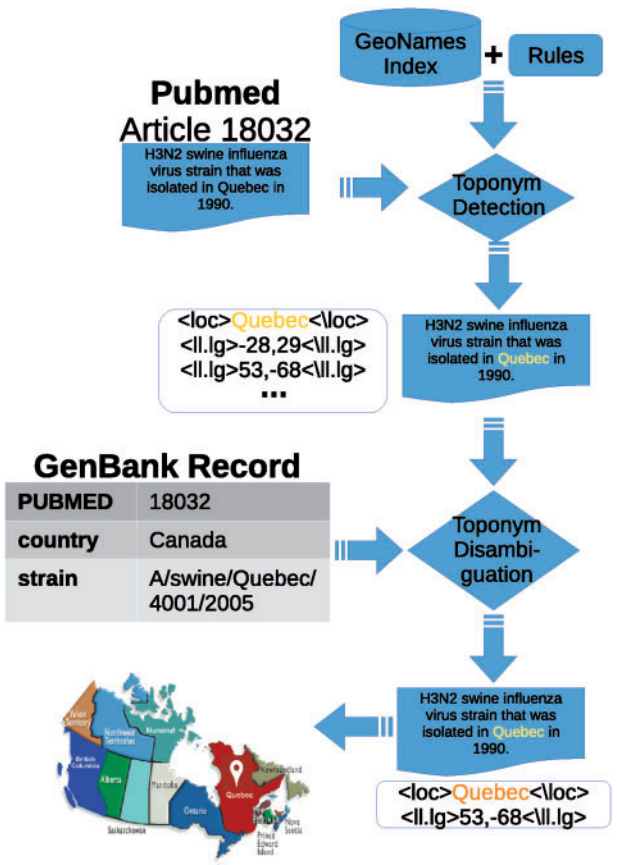

Fig. 1. Overview of the toponym resolver pipeline

Roberts *et al.* (2010) and Speriosu (2013) generalized this idea by modeling all discriminant words in the context of a toponym. Intuitively, if *Paris* is found in the context of words like *tourists, subway* and *restaurants*, it is more likely to be a tourism advertisement referring to the French capital rather than to the homonym Canadian city. To validate this hypothesis in his thesis, Speriosu ran several experiments on two types of corpora, newspaper and historical texts. The authors evaluated several resolvers. For the newspapers, the best resolver exhibited a precision of 0.84 P for the Disambiguation Only and 0.65 F-score for End-to-End evaluation. For the historical texts, the best resolver achieved a 0.87 P for the Disambiguation Only (an End-to-End evaluation was not performed).

2.1.2 Disambiguation driven by physical properties

The second approach in the literature exploits the physical properties of toponyms for disambiguation. Thus, all attributes describing geospatial locations, as well as their relations on earth, are modeled and used for disambiguation. This theme is behind several popular heuristics summarized by Leidner (2007). A first heuristic using attributes of the geospatial locations resolves the ambiguity between toponyms by choosing the most important location. The importance of a location is often computed by taking the place with the largest population. A second heuristic is based on the hypothesis that a document will usually refer to locations in a limited geographic area, so it chooses locations that are close to each other. For instance, if a document mentions the three ambiguous toponyms *Dallas*, *Paris* and *Houston*, this heuristic will assign the adjacent cities all located in Texas to these toponyms. In this article, we refer to the first heuristic as *Population heuristic* and to the second one as *Distance heuristic*.

2.1.3 Learning toponym disambiguation

The previous two approaches to disambiguation, driven by lexical content and physical properties, respectively, are not mutually exclusive and may be combined through the use of machine learning (ML) methods. Recent work, (Santos *et al.*, 2014), evaluated 58 features encoding the lexical context and the geospatial attributes of toponyms when performing toponym resolution. Their results were modest but proved that the combination of different features performs better than a single heuristic. Since their system used a different database of geospatial locations than the one they used to annotate the gold standard, the authors chose a particular metric that cannot be compared with the precision/recall metric commonly used in information extraction.

2.2 Disambiguation helped by metadata

In a similar approach than the one proposed in this study, Verspoor *et al.* (2012) successfully used metadata about a protein sequence cataloged in the Protein DataBank to filter candidate protein residues detected in corresponding PubMed abstracts. Document-related metadata describes the content in the document and the context where the document was produced. Document-related metadata can be of various kinds including predominant concepts in the document, the author or his/her institution. In their publication Zhang and Gelernter (2014) disambiguated toponyms in Twitter data using an SVM model with features such as population and alternative names from the GeoNames database along with Twitter metadata features such as user location. They combined the baseline features for toponym candidates with features indicating whether the user location overlaps with the candidates on country-level and state-level. The authors report precision, recall and F-score of 0.82 P, 0.79 R and 0.81 F-score respectively for Disambiguation Only.

2.3 Toponym resolution evaluation

Although significant progress has been made in the last decade on toponym resolution, it is still difficult to precisely determine the current state-of-the-art in performance achieved (Leidner and Lieberman, 2011). Tobin *et al.* (2010) suggest several reasons for this. The first difficulty is the lack of common frameworks for thorough comparisons of toponym disambiguation. The technology is recent and only a few standard corpora currently exist. We can cite for example TR-CoNLL (Leidner, 2007) and SpatialML (Mani *et al.*, 2008), two corpora from newspapers, or from a subpart of the wikipedia (Santos *et al.*, 2014).

The second difficulty arises from non-standard practices of the gazetteers (geospatial dictionary) for assigning coordinates to place names. The level of coverage may differ considerably from one resource to another and the latitude and longitude coordinates for the same location may vary as well, resulting in unjustified penalties when scoring the systems. As mentioned by Buscaldi (2011) *Cambridge* has only two entries in Wordnet, 38 in Yahoo! Geoplanet, and 40 in GeoNames. Even if a system selects the correct entry for this city based on its own set of resources, it may still assign slightly different coordinates from the gold standard and get unduly penalized.

Lastly, the large variation in corpora used for evaluating the different heuristics for toponym disambiguation, resulting from domain-based differences, also makes it difficult to determine the best heuristic for this task (Mani *et al.*, 2008). The difference in performance between the various heuristics applied may be a result of the different domain-specific assumptions implicitly shared between authors and readers but not known by toponym resolvers. For example when reading a local newspaper both authors and readers tacitly agree that all places are located in the same region, whereas as confirmed in Speriosu (2013), places cited in the major newspapers review important events occurring in the world and tend to be capitals or major cities distant from each other.

2.4 Toponym resolution on biomedical domain

To date, few publications have focused their studies on toponym resolution for biomedical documents. The first attempt was made in the work published by Turton (2008). Articles in PubMed Central are indexed according to a selected list of Medical Subject Heading (MeSH) terms. The list of MeSH terms contains important geospatial locations that have been found to occur in the abstracts by human annotators. Using a rule-based tagger, the author resolved toponyms in 1871 PubMed abstracts. To test the system, the author evaluated its ability to find in the abstract all mentions of the geospatial locations MeSH terms used to tag the document. The resolver performed relatively well since 70% of the geospatial locations MeSH terms occurring in the abstracts were correctly identified. The authors did not report any scores for End-to-End performance or for Disambiguation Only.

More recent attention has been focused on toponym resolution for microbial ecology. Tamames and de Lorenzo (2010) describe a system dedicated to the search of the precise locations of bacterial habitats. The authors evaluated their results on an internally developed corpus including 50 full-text articles and 200 abstracts. The authors performed toponym detection by matching entries of GeoNames and Google Maps with noun phrases in the corpus which contained capitalized words. They then trained a classifier to

distinguish between environmental or experiment-associated sentences and used the classifier to remove all putative locations appearing in experiment-associated sentences. The authors performed toponym disambiguation by searching for location mentions in the lexical context of an ambiguous toponym. When no other place name could be found in the context of an ambiguous toponym, the authors addressed the ambiguity by querying the GeoNames API that returned the possible locations sorted by relevance. The scores reported for the resolution in the full-text articles were good with an End-to-End system achieving 0.86 R and 0.92 P (the authors did not report scores for the toponyms Disambiguation Only). A challenge on a similar task was organized during the BioNLP Shared task 2011 (Bossy *et al.*, 2011). The goal of the task was to identify bacteria and the type of their habitat. Thus, the results of the challenge cannot be directly compared to our results since the corpus was a set of web pages for non-scientists, a corpus which differs largely from the phylogeography literature.

3 System architecture

In this section we present our toponym resolver for phylogeographic documents (Fig. 1). Our system is composed of two modules applied sequentially. The first module attempts to identify all names of places in a document. The second module addresses the disambiguation of each toponym identified by the first module; it takes as input all the names of places and assigns a unique pair of coordinates to each of them.

3.1 Toponym detection

GeoNames is a collaborative database of geospatial locations and is available at http://www.geonames.org/. We downloaded the version of the database available in the cited website and installed it locally. The installed database contains more than 10 millions names of places along with their alternative names, which include most of common transliterations for non-English names of locations. All entries are provided with their unique coordinates. During our initial set of experiments, we noted that the names of some countries or their common designations such as *South Korea* were missing in our version of GeoNames. To ensure the comprehensiveness of names of countries, we decided to add into our local GeoNames database all the names and common alternative names of all countries found in the socrata.com dataset (https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram).

To detect toponyms in documents, we built a rule-based module relying primarily on the GeoNames database and the Socrata dataset. Our module detects toponyms in documents by matching entries of GeoNames with the phrases in the documents. Since location names are of arbitrary length, each sequence of tokens in a document should, theoretically, be compared against all entries of GeoNames to find a match, leading to a problem of exponential comparisons. Due to the considerably large size of GeoNames, we built an index to perform more efficient look-ups. Our system splits all entries of GeoNames into contiguous sequences of tokens and then clusters them according to the first two letters of their first token. Each cluster is then associated with a variant of a Patricia tree, where edges are associated with a unique token and all nodes can either be accepting or non-accepting. Each token sequence representing an entry of GeoNames is modeled as a distinct branch of a tree, with its last node accepting. In Figure 2, we provide a concrete example for three locations: Abar, Abar abu Hileiqt, Abar abu Hirga. The creation of the index took 7.19 min on an Intel Xeon

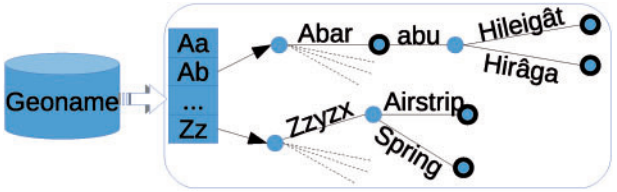

Fig. 2. Patricia tree used to index the GeoNames database. Black nodes are accepting nodes

2.60 Hz processor with 9 GB of memory and each document was parsed in 6.66 cs on average. The complexity of the search algorithm is in O(log n).

We created a black-list and a set of rules to remove noisy entries and common names found in GeoNames which would yield countless false positives (FPs) (such as *A, How, Although, Gene, Body*). Our black-list contains 1242 entries and is available at http://diego.asu.edu/downloads. We created the list after a manual examination of a development corpus composed of 28 documents from PubMed. We describe the development corpus in (Tahsin *et al.*, 2014). In addition to the black-list, we used a set of linguistic rules to filter out acronyms, names of people, names of organizations, and adjectival uses of the toponyms based on their immediate lexical context and grammatical properties. We kept only common acronyms for countries such as *USA* or *UK* and the postal codes of Canadian provinces and American states, which are very frequent in our corpus. To apply the rules, we preprocessed our documents using two external NLP tools: the ANNIE sentence splitter of the GATE pipeline (Available at https://gate.ac.uk/) to segment the sentences and the Genia tagger for POS tagging and chunking the phrases (Available at http://www.nactem.ac.uk/tsujii/GENIA/tagger/). Once a toponym is detected, an XML tag is substituted for it in the document and all possible coordinates for this toponym are inserted for further disambiguation.

3.2 Toponym disambiguation

Our toponym disambiguator is modular and can make use of various heuristics. It takes as input a document where all toponyms have been detected and assigned possible coordinates in latitude and longitude. It outputs a document where all toponyms are paired with their most likely coordinates. To perform this disambiguation, we have implemented two heuristics proven to be effective in previous studies (Leidner, 2007) - the population heuristic and the distance heuristic. In addition we have developed a novel heuristic that we call here the 'metadata heuristic' to further improve this process.

3.2.1 Population heuristic

This heuristic is computed based on the population field of the GeoNames database. For an ambiguous toponym, this heuristic will always choose the candidate which has the largest population. Thus *Paris* in France will always be preferred over *Paris* in Texas or Ontario. One major limitation of this approach is the presence of incorrect population sizes for some toponyms in the database. For example, in Geonames the population of the continent *Africa* is 0, and therefore, for the place *Africa*, this heuristic always chooses Madhia a small city of 45 thousand people in Tunisia, which is also called Africa.

3.2.2 Distance heuristic

This heuristic searches to minimize the surface of the polygon formed by linking all coordinates of the considered candidates for

all ambiguous toponyms present in a document. The exact solution could have been computed with linear optimization, but the number of candidates for an ambiguous toponym can be very high, e.g. in our database *San Jose* has 676 possible coordinates. We estimated a good candidate by implementing a search heuristic initially proposed by Speriosu (2013) to avoid issues with computational complexity. For each toponym T_i, the heuristic selects a candidate A_i which is at the minimum distance from all candidates of all other toponyms. The resulting choice, although good, is not necessarily the best one since the optimal solution may prefer T_i to a candidate B_i which is farthest away from A_i but minimizes the global distance between all best candidates selected for the other toponyms.

3.2.3 Metadata heuristic

This heuristic uses geospatial metadata in GenBank records to improve toponym resolution in related articles. The first step in this process extracts the most specific location for each GenBank record linked to an article by integrating geospatial information from different fields within the record. For instance the 'country' field in the record *JF340082* contains the geospatial metadata *Austria* while the 'strain' field contains the location *Vienna* embedded within the strain name *A/Vienna/25/2007*; when integrated together they produce a more specific location *Vienna, Austria*. In (Tahsin *et al.*, 2014), we provide a detailed description of the process through which the locations are extracted and combined. This extraction process involves the use of the administrative codes of locations recorded in GeoNames. States and provinces in GeoNames are 1st-level administrative divisions (ADM1), while counties, municipalities and towns have a different administrative code (ADM2 and below). The metadata heuristic largely depends on the specificity of the location extracted from these records and can be divided into the following four cases:

- Rule 1: If one of the GenBank records related to an article contains only the name of a country (e.g. *France*), then for every toponym which includes a geo-coordinate within this country (e.g. *Paris*), the system restricts the possible set of geo-coordinates for the toponym to those within this country. Toponyms that do not have a geo-coordinate within this country are not affected.
- Rule 2: If one of the records related to an article contains an ADM1-level location along with the name of a country e.g. (*VA, USA* or *Peterburg, VA, USA*), then for every toponym which includes a geo-coordinate within this specific administrative division in this specific country (e.g. *Petersburg*), the system selects a subset of geo-coordinates for the toponym which are present within this region. Toponyms that do not have a geo-coordinate within this region are not affected.
- Rule 3: If one of the records related to an article contains a location more specific than an ADM1-level location along with the name of the country in which it resides but does not mention its parent ADM1-level location (e.g. *Shantou, China*), then the system checks whether the article mentions any of its possible parent ADM1-level locations (e.g. *Guangdong* and *Fujian*). If only one of the possible parent ADM1-level locations is found, then that location is assumed to be the correct parent ADM1 and the heuristic in Rule 2 is applied. Otherwise the Rule 1 heuristic is applied.
- Rule 4: If one of the records related to an article contains a single location mention with no country information, then the system checks whether any of the countries in which the location can be possibly found is present in the content of the article. If only one of the possible countries is detected in the article, then the location is deemed to be present in this country and depending on whether the location is an ADM1-level geospatial entity within this country or more specific than ADM1-level, either the Rule 2 heuristic or the Rule 3 heuristic is applied. If no such country is found, or more than one possible country is found, then this heuristic cannot be applied.

Our heuristic holds the 'one-sense per toponym' hypothesis, namely, all occurrences of an ambiguous toponym found in an article have the same coordinates. For instance, if the virus in one of the GenBank records related to an article is known to have been collected from Shantou city in the Guangdong Province of China, we assume that any mention of *Shantou* within the article refers to this specific location. In addition we also applied a separate rule to identify all continents; when one of the candidates for a toponym was a continent, our system always chose the coordinates of the continent.

We derived these rules based on the observation that most of the toponym mentions in PubMed articles linked to a set of GenBank records refer to either the location of collection of the GenBank sequences or nearby regions. Therefore, for each toponym found in an article, the rules attempt to select the candidates present in the countries and/or ADM1-level locations mentioned in the GenBank records linked to the article.

The metadata heuristic is only applicable to articles related to at least one GenBank record with geospatial metadata. Moreover, in such articles, it can only be applied to toponyms having at least one candidate that shares its ADM1 code and/or country code with the geospatial metadata of at least one of the related GenBank records. Since the metadata heuristic, by itself, cannot be used to disambiguate all the toponyms detected in an article, we used the population heuristic to complement this task. For each ambiguous toponym that the metadata heuristic failed to solve, we chose the candidates with the highest population. In cases where more than one candidate was found to have the same population, we randomly selected one of them. For evaluation purposes, we gave preference to the population heuristic instead of the distance heuristic to complement the metadata heuristic since our results indicated it to be faster and more accurate. However, adapted heuristics can be developed based on the analysis of our results and replace or complement the population heuristic to further improve the disambiguation process.

4 Corpus and metrics

4.1 Corpus

We created a corpus dedicated to phylogeography to carry out our evaluation. We downloaded 102 949 GenBank records that were linked to NCBI taxonomy id 197911 for influenza A and for which a PubMed ID was listed. The exact request was http://www.ncbi.nlm.nih.gov/nuccore/?term=txid197911[Organism:exp]. The downloaded records were associated with 1424 distinct PubMed articles and 598 of them had links to an open access journal article in PubMed Central (PMC). We randomly sampled 60 articles from this set of 598 articles for manual annotation. These articles are associated with 5730 distinct GenBank records from our original set and represent a total of 500 000 tokens. About 99.9% of these records (5726 records) contained some form of geospatial metadata. However, the metadata were more specific than ADM1-level for only 25.5% of the records (1461 records). In prior work, Scotch *et al.* (2011) characterized ADM1-level geospatial information to be imprecise for local phylogeography studies. Therefore, 74.5% of these records (4269 records) may lack sufficient geospatial metadata

to meet the needs of phylogeography requiring researchers to search through related PubMed articles for more specific information

To perform the annotation, we manually downloaded the PDF versions of the articles and converted them to text files using the freely available tool, pdf-to-text (Available at http://www.foolabs.com/xpdf/download.html). Our toponym detector was applied on the text files and we formatted the output to be compatible with the Brat annotator (Available at http://brat.nlplab.org/). We manually corrected and disambiguated the toponyms using GeoNames. We annotated toponyms in the body of the document, the tables and the captions. We removed contents that would not contain phylogeography-related toponyms, such as the names of the authors, acknowledgments and references. Although this was done manually for our experiments, these areas could have been detected automatically with high accuracy [>0.90 F1 score (Tkaczyk *et al.*, 2014)]. In cases where a toponym couldn't be found in GeoNames, we searched Google Maps and Wikipedia. If the toponym was still not found, we set its coordinates to a special value *NA*. Prior to beginning annotation, we developed a set of annotation guidelines after discussion among four annotators familiar with the biological domain. The resulting guidelines and the gold standard are available at http://diego.asu.edu/downloads.

The gold standard contains a total of 379 distinct toponyms for a total of 1881 occurrences. Two hundred sixty-four of these toponyms are present in only one document (a document may include multiple occurrences). The average number of occurrences for a toponym is four with the most cited toponym, *China*, having a total of 209 occurrences in our corpus. The average ambiguity is about 19 candidates per toponym which is close to the average ambiguity found in existing corpora Speriosu (2013). The location *San Jose* was found to have the maximum ambiguity of 676. We found that the distribution of toponyms across the different major sections of a document, such as title, document body, figures labels and table content, to be fairly even; multiple toponyms were in any of these sections.

For this study, we prioritized recall (i.e. minimizing the number of false negative toponyms) over precision (i.e. minimizing the number of false positive toponyms). One reason for this choice was to reduce the work of the annotators and to speed up the creation of our gold standard corpus. During the annotation process, we found that removing false positives in the document was easier than searching for the corresponding entry in our database for a missed toponym. Another reason was that the overall goal of the larger project is to complete our local GenBank database records with the correct locations of the viruses. It is therefore more important to find all locations in a document since a location missed cannot be associated with the corresponding virus at a later stage whereas a false positive can always be removed when pairing a virus with a location.

4.2 Metrics

The standard metrics of precision, recall and F-measure can be used to report the performance when the gold standard and the system are aligned on the same knowledge base. For our experiments, we report our results by using two common variations of these metrics: strict and overlapping measures. In the strict measure, the system annotations are considered matching with the gold standard annotations if they hit the same spans of text; whereas in the overlapping measure, both annotations are matching when they share a common span of text.

We computed the Precision and Recall for toponym detection with the standard equations (For toponym detection, Precision=$\frac{TP}{TP+FP}$ and Recall=$\frac{TP}{TP+FN}$, where TP is the number of toponyms correctly identified by the system in the corpus, FP the number of phrases incorrectly identified as toponyms by the system, and FN the number of toponym not identified by the system.). The precision and recall for toponym disambiguation is given by slightly modified equations. The precision P_{ds} of the toponym disambiguation is given by the equation 1, where T_{CD} is the number of toponyms correctly identified **and** disambiguated by the system in the corpus and T_{ID} is the number of toponyms incorrectly identified **or** disambiguated in the corpus.

$$P_{ds} = \frac{T_{CD}}{T_{CD} + T_{ID}} \quad (1)$$

The recall R_{ds} of the toponym disambiguation is computed by the fraction 2, where T_N is the total number of toponyms in the corpus.

$$R_{dt} = \frac{T_{CD}}{T_N} \quad (2)$$

Since our system and the gold corpus annotations are both aligned on GeoNames, toponyms correctly identified are known by a simple match between the place IDs retrieved by the system and those annotated by the experts. Not matching on coordinates avoids the problem of having the system and the gold standard denoting two different toponyms but referring to the same coordinates; for instance, a city and its state may have the same geo-coordinates in GeoNames but they refer to different locations and hence will have different place IDs.

4.3 Inter-annotator agreement

We computed the inter-annotator agreement on 16 articles selected randomly within our gold standard corpus. Two different annotators annotated the 16 articles independently. Since our task is a named-entity recognition task, we follow the recommendations of Hripcsak and Rothschild (2005) and used P and R to estimate the inter-annotator rate. The inter-annotator agreement rate on the toponym detection was 0.97 P and 0.98 R which indicates a good agreement between the annotators. On a total number of 509 toponyms, annotators disagreed on 30 occurrences (18 FPs, 12 FNs in strict evaluation). Twenty occurrences were found to be annotation errors. Of these 20, 11 were toponyms appearing in the document's metadata zones which should have been removed by the annotator according to our guidelines, and nine occurrences were correct toponyms but missed by the annotators. For the other 10 disagreements, six occurrences were disagreements on the boundaries of location names (for ex. *Kaduna State* or *Republic of Laos*), and four occurrences were caused by different interpretations of the guidelines with names like *Eurasia* or *Federal Capital Territory*, considered as names of places by one annotator but not by the other. The inter-annotator agreement on both toponym detection and disambiguation was 0.96 P and 0.98 R. Cases where the boundaries of the toponyms were correctly set but annotators disagreed on their coordinates accounted for three occurrences caused by variation in the interpretation of Korea (one annotator chose the north side and the other the south) and Central Europe (which is not an entry in GeoNames and the annotators found two different coordinates on the web).

5 Evaluation

5.1 Toponym detection performances

During the first experiment, we evaluated the performance of our toponym detector only. For this experiment, we did not attempt to

disambiguate the toponyms detected by our module. The results show a good coverage of the GeoNames database over our corpus with 0.90 R (When our toponym detector is run without applying the black-list and the rules defined to filter out false positive in section 3.1, 99.1% of the toponyms in our corpus are found. This confirms the good coverage of GeoNames on our corpus.) in Table 1 (first line). As a result of a high recall, the precision achieved was only 0.599 P.

To understand the behavior of our toponym detector, we randomly selected 10 documents and analyzed by hand the causes of the errors. The more frequent FPs were a result of acronyms and abbreviations being recognized by the system as names of places; they account for 55% of the system's errors. However the previous percentage misrepresents their importance since one single document contained 126 occurrences of the same error, i.e. 52% of the total number of the errors. These errors occurred in a table within the document where *Bei* abbreviates *Beijing* in strains. The most serious cause of FPs were the ambiguities between toponyms and the names of people/organizations/animals or simply common words in the English dictionary. We have encountered, for example, *Andrew [Caton], [Life] Science, Condor, or Levels* as ambiguous names. These ambiguities account for 13.3% of the errors. Another important cause of errors, representing 12.9% of the system's errors, was the incorrect preprocessing of the documents. Incorrect annotations computed during the preprocessing of the document caused the system to wrongly trigger some rules and discard (or keep) unexpected phrases. For instance, incorrect segmentation of sentences during the PDF conversion caused the detector to take *Bar* as a city and not *Bar Harbor*. Other examples include several toponyms for which their part of speech were incorrectly annotated by our system like *Hong Kong* in *Hong Kong flu*, annotated as nouns and not adjectives, or *Beijing* being tagged as a verb. These latter examples should have been removed and kept, respectively, since our rules accept only toponyms in nominal position. 10.4% of the errors were caused by technical terms used in this domain that can also be the names of toponyms. Biological terms like *Sigma* or *Tris* and names of genes/proteins represented the majority of this type of errors.

It is apparent from the Table 1 and the analysis of the previous errors that the major limitation of our toponym resolver comes from the toponym detector module. Alternative solutions to our rule-based module exist. We have recently tested a module using Conditional Random Fields (CRF) for recognizing the toponyms. CRF, a sequence labeler, is currently one of the best statistical machine learning frameworks for named entity recognition (McCallum and Li, 2013). We trained the CRF on our development corpus composed of 28 full-text articles extracted from PubMed. However, with our current setting, we observed low performance on our test corpus with 0.539 F-score. A possible explanation for this under-performance might be the difference between the two distinct guidelines used to annotate the toponyms in the development corpus and the ones we agreed on for our test corpus. The guidelines used to annotate the development corpus by Tahsin *et al*. (2014) is different in various aspects from the guidelines defined specifically for our toponym resolution task. In our development corpus we considered as names of locations and annotated toponyms in adjectival positions, such as *Fort Morgan virus*, as well as toponyms nested within strains. These cases were not annotated as toponyms in the test corpus used for the present work. The high frequency of such cases in our corpora may have caused a bias during the learning of the prediction model explaining the low precision of only 0.60 P. We are currently modifying the annotation of our development corpus according to the guidelines of the test corpus and will retrain our CRF.

Table 1. Toponym resolution on phylogeographic documents using various heuristics

	P	*R*	*F*
1. Detection only	0.58/0.599	0.876/0.904	0.698/0.72
Disambiguation only			
2. Population	0.754	1.0	0.86
3. Distance	0.625	1.0	0.769
4. Metadata + population	0.886	1.0	0.939
End-to-End (both detection and disambiguation)			
5. Population	0.474/0.492	0.854/0.887	0.610/0.633
6. Distance	0.394/0.412	0.83/0.868	0.534/0.559
7. Metadata + population	0.528/0.547	0.867/0.897	0.657/0.679

Scores are given in precision, recall and F-measure where exact/overlapping toponyms are considered respectively.

5.2 Toponym disambiguation performance

In a second experiment, we evaluated the performance of our toponym disambiguator only. Here, we did not perform toponym detection and we gave the boundaries of all named entities in the documents to the toponym disambiguator and only analyzed for disambiguation. We used the two heuristics, *Population* and *Distance heuristics*, separately to compare their respective precisions. In Table 1 (second and third lines), we show that there is a clear advantage of the population heuristic over the distance heuristic with a difference of 0.13 points of precision.

As for the toponym detection, we analyzed the errors made by the toponym disambiguator when using the population heuristic on the same 10 documents. Causes of errors can be classified into three main categories. The first category accounts for 44% of the errors and are directly or indirectly attributable to the imperfection of GeoNames. Only one place name was not referenced in GeoNames (*Gurjev* an alternative name for Astrakhan). One common error was made on 12 occurrences of the European continent; GeoNames references various coordinates for denoting its centroid and the disambiguator chose centroid coordinates that were different from the gold standard. The last source of error in this category was missing population information for some places in GeoNames such as the African continent, causing the disambiguator to choose another toponym with a higher population. The second category included errors caused by the simplicity of the heuristic used for disambiguation, accounting for a total of 42% of the errors. Fifteen errors were caused by an incorrect phrase segmentation made during the PDF conversion. Special algorithms for revising the segmentation may correct the errors. Four errors resulted from the system being unable to distinguish between a populated city and an ADM1-level place. When using the *population heuristic*, the disambiguator always chooses the most populated place. Although this choice is, on our corpus, most often the correct one, for some cases it is not. For example, when a company name is specified in the text, the city following the name of the company's headquarters should be favored over the administrative place. Only one error was an ambiguous name of a city incorrectly resolved by choosing a homonymous city with the highest population: *Cleveland* in UK is chosen by our algorithm when the city in Ohio was expected. The last category of errors, 14% in total, is due to misspelled toponyms in articles (e.g. *Oya State* for *Oyo State*) and errors made by the annotators which will be corrected during a second pass of annotation.

When we applied the metadata heuristic along with the population heuristic, many of the errors produced by the system were eliminated, leading to a 0.26 point increase in precision Table 1 (fourth line). For instance the system was able to correctly identify the coordinates of *Thai Binh* in Vietnam since it now gave preference to the ADM1-level location in Vietnam based on existing geospatial metadata instead of selecting one of the nine possible coordinates for this location in GeoNames, all of which had a recorded population of 0. A significant increase in performance was also achieved by always selecting the toponym candidate with the continent feature code in GeoNames in case of continent names. As a result of this filter, the system was able to correctly identify the GeoNames IDs for frequently occurring continent names such as Africa, North America and Asia.

For the metadata heuristic, error analysis of the 10 documents revealed that the majority of the errors were caused by misspelled toponyms in articles, annotation errors, and some minor deviations from the annotation guidelines. They account for 50% of the total number of errors (11 instances) in these documents. In total, 41% of the errors (nine instances) were caused by limitations of the algorithm, with 9% (two instances) being a direct result of the metadata heuristic. For instance, for the state abbreviation *MO*, the system chose a location in Vietnam based on the metadata heuristic instead of the state of Missouri in USA; the population heuristic by itself would have chosen the correct candidate. The remaining 9% of the errors (two instances) were a result of GeoNames imperfection.

The major limitation of the metadata heuristic is its limited applicability; it can only be applied to publications associated with GenBank records that contain geospatial metadata. However, in our corpus of 60 PMC articles, 99.8% of the 5730 GenBank records associated with the articles contained some form of geospatial metadata and 30% (531 toponyms) of the toponyms present in our corpus were affected by this heuristic. Although we have tested our approach on a small subset of records related to articles on the influenza A virus, it can be applied to any PubMed article linked to GenBank records. Based on an analysis of 87 116 501 GenBank sequence files performed in Miller *et al*. (2009), about 30% of GenBank records are linked to at least one PubMed article. Under the reasonable assumption that these articles are on average linked to 100 GenBank records with 90% of these records containing some form of geospatial metadata, our heuristic can be applied to over 200 000 PubMed articles.

5.3 End-To-end system performance

The last part of the Table 1 (fifth, sixth and seventh lines) shows the results of the End-to-End toponym resolver making use of the three heuristics separately. In this experiment the toponym detector and resolver are applied sequentially without human intervention.

The performance of our End-to-End resolver is very competitive with the performance of toponym resolvers working with lexical contexts. However, our results fall short of those reported by Tamames and de Lorenzo (2010) whose End-to-End system scores on similar corpora 0.86 R and 0.92 P. A possible explanation for this might be that Tamames and de Lorenzo (2010) made the choice to work only with XML conversion of the full articles. This choice is too restrictive for the overall goal of our project. XML conversions are clean texts compared to texts converted from PDF. By working on clean texts their system avoids the bulk of the errors made by our system during the detection of toponyms and caused by the poor quality of the conversion of the documents, either directly or indirectly. Errors directly caused by conversion errors may be the result of unformatted tables and the insertion of article metadata, such as author names and affiliations, in the content of the document. Indirect consequences are imposed by the challenges faced when pre-processing noisy documents leading to wrong sentence segmentations and POS tagging errors. We are currently implementing the algorithm proposed by Tamames and de Lorenzo (2010) and will evaluate it on our corpus to confirm this hypothesis.

6 Conclusion and future work

The goal of this study was to determine the performance of an automatic toponym resolver on phylogeography articles. Our toponym resolver makes use of an innovative heuristic relying on metadata available in GenBank to solve geo-geo ambiguities. When applying the metadata heuristic, our system extracts and compiles all geospatial information available in the fields of the GenBank records to formulate a hypothesis about the area on earth where the virus was located. This helps to restrict possible candidates during the disambiguation of names of locations in texts and improves the accuracy of the inference.

Since there was no annotated corpus available for evaluating our system within the domain of phylogeography, we utilized full-text PubMed Central articles linked to GenBank records to develop a corpus of our own. Our toponym resolver achieved an 0.69 F-score when performing End-to-End resolution. When only performing disambiguation of the toponyms annotated in the gold standard corpus, it had a precision of 0.88 P. The metadata heuristic proposed in this article led to a 26% increase in precision over using the standard population or distance heuristics.

We are currently using the toponym resolution approach presented in this article to link GenBank records to the latitude and longitude coordinates of the most specific location of collection available for them using data from the GenBank record fields and related articles. Our preliminary results on this task are encouraging but indicate a need for improvement of the toponym detection.

As future work, we are aiming to replace our rule-based module for toponym detection by exploring a method to semi-automatically acquire the examples required to train a CRF using the geospatial metadata of GenBank records. By using the geospatial information reported in the *country* and *location* fields it is possible to automatically extract the sentences where geospatial information is found in their PubMed articles. This substantially large set of automatically extracted sentences can then be used as positive examples to train our CRF and improve its performance when detecting the toponyms.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R56AI102559 to M. Scotch and G. Gonzalez.

Conflict of Interest: none declared.

References

Adams,B. and McKenzie,G. (2013) Inferring thematic places from spatially referenced natural language descriptions. In: Sui,D. *et al*. (eds) *Crowdsourcing Geographic Knowledge*. Springer, The Netherlands, pp. 201–221.

Avise,J.C. (2000) *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Benson,D.A. *et al*. (2011) Genbank. *Nucleic Acids Res.*, **39**, 32–37.

Bossy,R. *et al*. (2011) Bionlp shared task 2011—bacteria biotope. In: *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

Buscaldi,D. (2011) Approaches to disambiguating toponyms. *SIGSPATIAL Special*, **3**, 16–19.

Hripcsak,G. and Rothschild,A.S. (2005) Agreement, the f-measure, and reliability in information retrieval. *JAMIA*, **12**, 296–298.

Leidner,J.L. (2007) *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh.

Leidner,J.L. and Lieberman,M.D. (2011) Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL*, **3**, 5–11.

Mani,I. *et al*. (2008) Spatialml: Annotation scheme, corpora, and tools. In: Calzolari,N. *et al*. (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco.

McCallum, A. and Li,W. (2013) Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: *Proceedings of CoNLL-2013*, pp. 188–191.

Miller,H. *et al*. (2009) Genbank and pubmed: how connected are they? *BMC Res. Notes*, **2**, 101.

Piskorski,J. and Yangarber,R. (2013) Information extraction: past, present and future. In: Poibeau,T. *et al*. (eds) *Multi-source, multilingual information extraction and summarization, theory and applications of natural language processing*. Springer, Berlin, pp. 23–49.

Roberts,K. *et al*. (2010) Toponym disambiguation using events. In: *FLAIRS Conference'10*, p. 1.

Santos,J. *et al*. (2014) Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, Springer Netherlands, 1–18.

Scotch,M. *et al*. (2011) Enhancing phylogeography by improving geographical information from genbank. *J. Biomed. Inf.*, **44**, 44–47.

Speriosu,M.A. (2013) *Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision*. PhD thesis, University of Texas.

Tahsin,T. *et al*. (2014) Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses. *AMIA Jt. Summits Transl. Sci. Proc.*, **2014**, 102–111.

Tamames,J. and de Lorenzo,V. (2010) Envmine: a text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, **11**, 294.

Tkaczyk,D. *et al*. (2014) Cermine—automatic extraction of metadata and references from scientific literature. In: *Proceedings of 11th IAPR International Workshop on Document Analysis Systems*, pp. 217–221.

Tobin,R. *et al*. (2010) Evaluation of georeferencing. In: *Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10*, pp. 7:1–7:8.

Turton,I. (2008) A system for the automatic comparison of machine and human geocoded documents. In: *Proceedings of the 2nd International Workshop on Geographic Information Retrieval, GIR '08*, pp. 23–24.

Verspoor,K.M. *et al*. (2012) Text mining improves prediction of protein functional sites. *PLoS One*, **7**, e32171.

Zhang,W. and Gelernter,J. (2014) Geocoding location expressions in Twitter messages: A preference learning method. *J. Spatial Inf. Sci.*, **9**, 37–70.